PROTOCOLO DE ANÁLISIS DE SECUENCIAS GENÓMICAS DE *LISTERIA MONOCYTOGENES*

Programa Nacional de Vigilancia

Gianecini Ricardo Ariel & Cipolla Lucía

Instituto Nacional de Enfermedades Infecciosas (INEI) – ANLIS "Dr. Carlos G. Malbrán, Departamento de Bacteriología, Servicio de Bacteriología Especial

CIUDAD AUTÓNOMA DE BUENOS AIRES, MAYO DE 2024

ANLIS Dr. C. G. Malbrán. Instituto Nacional de Enfermedades Infecciosas. Servicio de Bacteriología Especial. Protocolo de Análisis de Secuencias Genómicas de *Listeria monocytogenes*. Ciudad Autónoma de Buenos Aires: ANLIS Dr. C. G. Malbrán, 2024. Disponible en: https://sgc.anlis.gob.ar/handle/123456789/2615

"Este recurso es el resultado del financiamiento otorgado por el Estado Nacional, por lo tanto, queda sujeto al cumplimiento de la Ley Nº 26.899 y la política de gestión del conocimiento de la ANLIS".

Protocolo de Análisis de Secuencias

Genómicas de *Listeria monocytogenes*

CABA, Mayo 2024

Inicio de vigencia desde: 01-05-2024

INDICE

1.- INTRODUCCION

Listeria monocytogenes es el agente causal de la listeriosis, una de las enfermedades de transmisión alimentaria más graves que afecta a humanos, con una mortalidad estimada de 20 a 30%. Esta infección suele aparecer en forma de casos esporádicos y en pequeños brotes. La vigilancia epidemiológica de los casos humanos y matrices alimentarias resulta fundamental para la identificación de asociaciones y sitios comunes de infección. Además, proporcionado información de relevancia para la elaboración temprana de medidas preventivas de control y protección. Actualmente, la secuenciación de genoma completo permite un elevado nivel de resolución y es considerada la técnica de elección para la vigilancia de *L. monocytogenes*. Esta posibilita la detección de brotes, evaluar la dinámica poblacional de aislamientos con distintos perfiles de virulencia y la identificación de reservorios o sitios comunes de infección. Además, la combinación de datos genómicos de diversas fuentes y epidemiológicos de distintas regiones geográfica permiten un conocimiento integral, el cual resulta fundamental para el manejo de esta patógeno, tanto para las autoridades de salud pública como para los actores de la industria alimenticia.

Este procedimiento aborda el estudio genómico de aislamientos de *L. monocytogenes* para su vigilancia en Salud Pública. Dicho procedimiento está basado en un algoritmo que utiliza línea de comando (UNIX) e incluye varias etapas de análisis según se describe a continuación

2.- DEFINICIONES Y ABREVIARUTAS

Reads: lecturas (fragmentos de nucleótidos obtenidos como producto de la secuenciación)

Short reads = Lecturas cortas: fragmentos cortos obtenidos utilizando plataformas de segunda generación (Ej., Illumina, Ion Torrent).

Raw data / raw reads = secuencias obtenidas como producto de la secuenciación cuya longitud dependerá de la plataforma y kits de secuenciación.

SGC: secuenciación de genoma completo

Symlink: acceso directo a las secuencias

Path: ruta. Es la forma de referenciar un archivo o directorio.

Fastq: archivo de texto que contienen los datos de la secuencia generada.

SNPs: Polimorfismos de nucleótido único

MLST: Tipificación por secuenciación de múltiples locus

3.- INSTRUCCIONES

Para llevar a cabo el siguiente procedimiento de análisis se deberán seguir todos los pasos detallados a continuación:

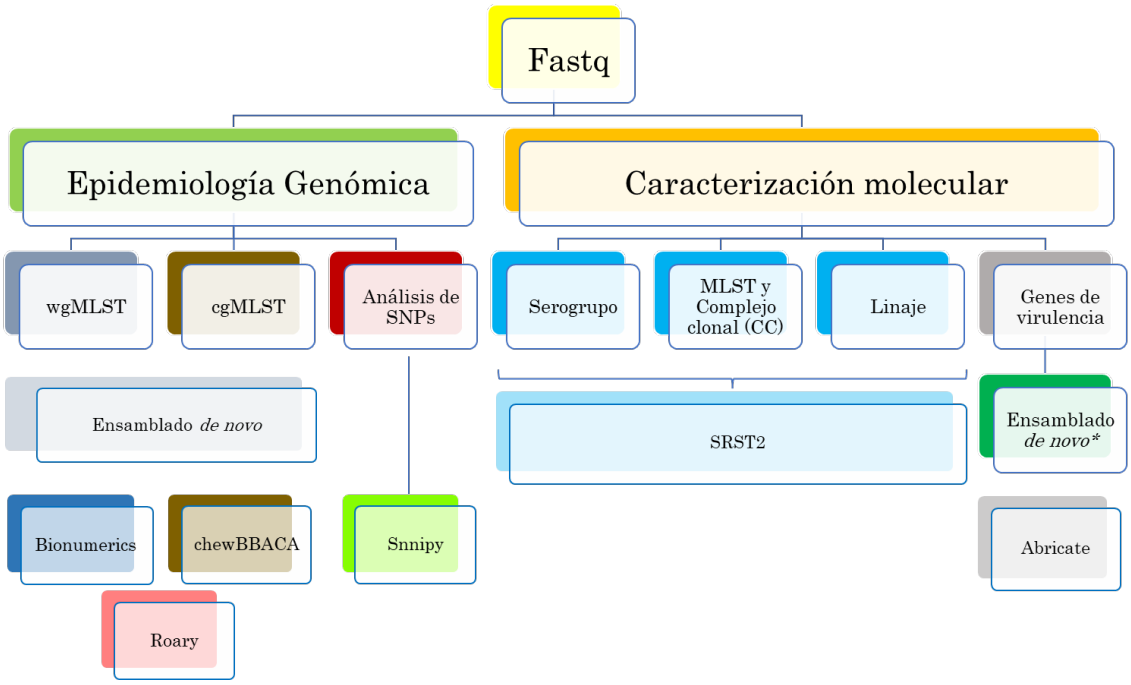
1) Para poder acceder a un servidor local o a un servidor regional para realizar el análisis se necesita establecer una conexión apropiada. Para ello se necesita la instalación de los programas WinSCP y Putty que son de uso libre y gratuito.

El WinSCP nos permite visualizar los archivos y resultados que se generan a partir del análisis, facilita la creación de directorios y la descarga de los mismos a nuestra computadora de trabajo. Putty es un programa que emula trabajar en línea de comando, a través de lo que se llaman clientes SSH, Telnet, rlogin, y TCP (esto dependerá de cada servidor) y permite correr comandos UNIX en el servidor al cual tengamos acceso.

El ingreso al servidor se hace a través de un usuario autorizado y contraseña que otorga resguardo de las tareas ya que solo se podrá trabajar en el espacio asignado a cada usuario.

NOTA: el acceso al servidor y formato de seq dependerá de la organización del servidor de uso.

1) Flujograma de trabajo



4.- PROCEDIMIENTO

4.1. Llamado de las secuencias a analizar

Para facilitar los análisis en el espacio de trabajo y no ocupar mayor espacio de lo deseado se recomienda trabajar con symlinks o accesos directos de las secuencias.

El comando "ln -s" da la indicación de generar el symlink y a continuación se detalla el path con la ubicación exacta de las secuencias originales.

Se utiliza el símbolo asterisco (*) para indicar todos los caracteres idénticos del nombre de las secuencias. Además, un espacio y un punto al final del comando indica que los symlinks sean creados en la ubicación actual del directorio.

ln -s /home/inei/secuencias/projects/BE/2309213/BE32* .

Muchas veces las secuencias cuando salen del secuenciador tienen nombres muy largos por lo que se pueden renombrar para facilitar su análisis.

Renombrar una secuencia:

mv BE-423_L001_R1_001.fastq.gz BE-423_1.fastq.gz

Renombrar varias secuencias:

for f in *_S*_L001_R1_001.fastq.gz; do mv $f

${f%_S*_L001_R1_001.fastq.gz}_1.fastq.gz; done

for f in *_S*_L001_R2_001.fastq.gz; do mv $f
${f%_S*_L001_R2_001.fastq.gz}_2.fastq.gz; done

Donde el símbolo asterisco (*) indica la parte variable del nombre
NOTA: depende de cómo se indique, el comando “mv” también se utiliza para mover archivos de un directorio a otro.

4.2. Calidad e identificación de las secuencias

Evaluar la calidad de las secuencias obtenidas mediante el programa FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Un *Phred score* >30 (Q30) es un indicador de buena calidad, ya que esto nos indica con una probabilidad >99,9% que el nucleótido leído es correcto. Las secuencias con una calidad mayor a Q30 para más del 80% de las bases serán utilizadas para los análisis.

mkdir fastqc
fastqc -t 4 -o fastqc *fastq.gz

Se obtienen dos tipos de archivos de salida, html y zip, para visualizar los resultados. Los parámetros mínimos que se deben considerar para poder avanzar con el análisis de las secuencias son:

- *Basis Statistics*
- *Per base sequence quality*
- *Per sequence quality scores*
- *Per base sequence content*
- *Per sequence GC content*

4.3. Identificación taxonómica de las secuencias

Realizar la identificación taxonómica de las secuencias mediante la herramienta Kraken2 (*Wood*, 2014). Esta herramienta utiliza un algoritmo basado en k-meros para determinar la taxonomía de las secuencias de ADN procedentes.

for f in *_1.fastq.gz ;do kraken2 -db /KRKDB2 --threads 16 --gzip-compressed --paired --report ${f%_1.fastq.gz}_kraken2.txt --use-names $f ${f%_1.fastq.gz}_2.fastq.gz; done

El output de elección será: *_kraken2.txt.

El formato de informe de muestra estándar está delimitado por tabulaciones con una línea por taxón y de izquierda a derecha, se leen los siguientes parámetros:

1. Percentage of fragments covered by the clade rooted at this taxon
2. Number of fragments covered by the clade rooted at this taxon
3. Number of fragments assigned directly to this taxon
4. A rank code, indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. Taxa that are not at any of these 10 ranks have a rank code that is formed by using the rank code of the closest ancestor rank with a number indicating the distance from that rank. E.g., "G2" is a rank code indicating a taxon is between genus and species and the grandparent taxon is at the genus rank.
5. NCBI taxonomic ID number
6. Indented scientific name

4.4. Identificación de serogrupo y tipificación por secuenciación de múltiples locus (MLST)

4.4a Descargar la base de genes y perfil de alelos de MLST y serogrupo, a través de BIGSdb-Lm (https://bigsdb.pasteur.fr/listeria/) o con línea de comandos.

Identificación de la base de interés

pubmlst_list --base-url https://bigsdb.pasteur.fr/api/db

Descarga del esquema seleccionado

mkdir mlst_scheme

pubmlst_download -s listeria -i 3 –outdir /path/OutputFolderName --base-url https://bigsdb.pasteur.fr/api/db

Opciones:
-s microorganismo
-i base de dato

4.4b Identificación de secuenciotipo (ST) de MLST y serogrupo mediante la herramienta *Short Read Sequence Typing for Bacterial Pathogens* (srst2) (*Inouye*, 2014)**.** Esta herramienta permite la tipificación molecular de bacterias mediante la estrategia de mapeo de las lecturas generadas.

Secuenciotipo
srst2 --input_pe *.fastq.gz --output test --log --mlst_db Listeria_monocytogenes.fasta --mlst_definitions listeria.txt --mlst_delimiter '_' --threads 16

Serogrupo
srst2 --input_pe *.fastq.gz --output test --log --mlst_db serotype.fasta --mlst_definitions profiles_sero.txt --mlst_delimiter '_' --threads 16

Para optimizar la capacidad del servidor y no ocupar mayor espacio es conveniente borrar los archivos intermedios:

rm *.pileup | rm *.bam | rm *.bt2
Utilizar el archivo de salida [outputprefix]__mlst__[db]__results.txt (ej. test_mlst_Listeria_monocytogenes_results.txt) para visualizar los resultados.

4.5. Evaluación de la relación genética de los aislamientos

La dinámica poblacional de los aislamientos se evaluará mediante la estrategia de tipificado basado en MLST de genoma central (cgMLST), pangenoma y el análisis de SNPs basado en referencia.

4.5.1. Determinación de cgMLST

La herramienta chewBBACA (*Silva*, 2018) se utilizará para la identificación del cgMLST, empleando la combinación de alelos de 1748 genes propuesto por el Instituto Pasteur de Francia como esquema (https://bigsdb.pasteur.fr/listeria/). Esta herramienta utiliza para la evaluación de los aislamientos el ensamblado de *novo* de los mismos.

4.5.1a Ensamblado *de novo* de los aislamientos

La herramienta Unicycler (*Antipov*, 2016) se utilizará para el ensamblado de *novo* de los aislamientos. Esta herramienta emplea el algoritmo de SPAdes para la generación de los ensamblados.

for f in *_1.fastq.gz ;do unicycler -1 $f -2 ${f%_1.fastq.gz}_2.fastq.gz -o assemblies/${f%_1.fastq.gz}_uni.out --verbosity 2 -t 24; done

Extracción de los archivos .fasta de los ensamblados

cd assemblies

for f in *.out/assembly.fasta; do mv $f ${f%.out/assembly.fasta}.fasta; done

Remoción de "_uni.fasta" de los nombres de los archivos

for f in *_uni.fasta; do mv $f ${f%_uni.fasta}.fasta; done

4.5.1b Control de calidad de los ensamblados

El control de calidad de los ensamblados se realizará mediante la herramienta Quast (https://github.com/ablab/quast).

quast.py *.fasta

Abrir el archivo report.txt para visualizar los resultados.

Criterio de aceptación de los ensamblados

- Longitud total del ensamblado: 3,0 ± 0,3 Mb
- Valor de N50 mayor de 30 kb
- Número total de contigs menor de 200

4.5.2. Identificación de cgMLST mediante chewBBACA

Este programa permite la creación de esquemas de MLST de genoma central (cg) y genoma completo (wg). Además, permite la adaptación de esquemas externos como PubMLST (https://bigsdb.pasteur.fr/) y Enterobase (https://enterobase.warwick.ac.uk/).

4.5.2a Descarga del esquema de 1748 genes de cgMLST

pubmlst_download -s listeria -i 3 –outdir /path/to/OutputFolderName --base-url https://bigsdb.pasteur.fr/api/db

4.5.2b Adaptación del esquema externo

chewBBACA.py PrepExternalSchema -i /path/to/ExternalSchema -o /path/to/OutputFolderName --ptf /path/to/ProdigalTrainingFile --cpu 4

4.5.2c Determinación del perfil alélicos de los genomas

chewBBACA.py AlleleCall -I /path/to/InputAssemblies -g /path/to/SchemaDirectory -o /path/to/OutputFolderName --cpu 4

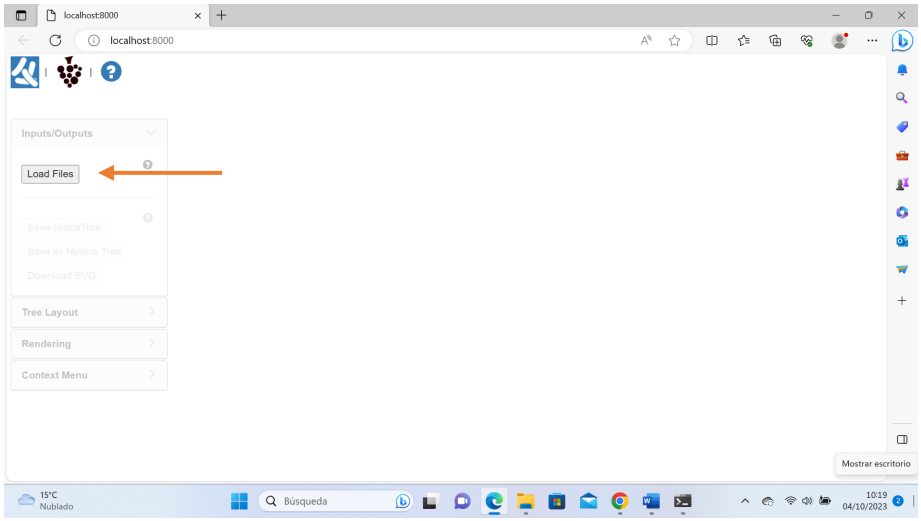
4.5.2d Determinación del grupo de alelos que constituyen el genoma central

chewBBACA.py ExtractCgMLST -I /path/to/results_alleles.tsv -o /path/to/OutputFolderName

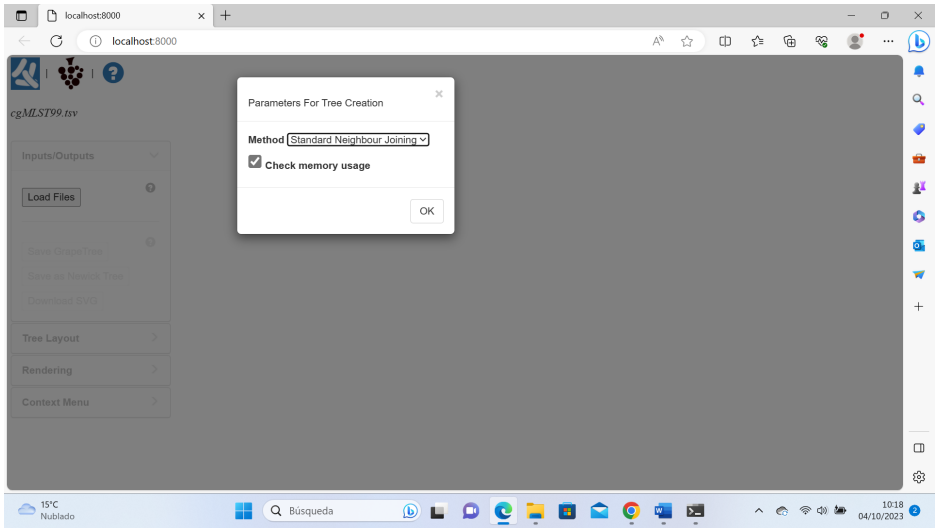
4.5.2e Visualización del esquema

Construir el dendrograma mediante el programa GrapeTree (*Zhou*, 2018). Utilizar el archivo cgMLST100.tsv como entrada e iniciar el nombre de la primera columna "Sample" con el símbolo "#" (ej. #Sample).

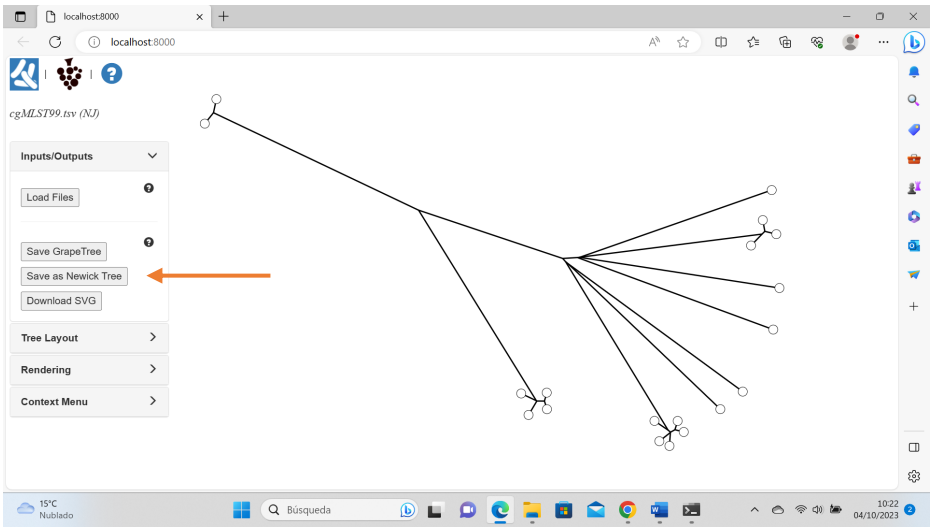
Cargar el archivo en "Load Files"



Seleccionar en los parámetros de creación del dendrograma "Standard Neighbour Joining"



Guardar el dendrograma en formato .newick mediante “Save as Newick Tree”. Este formato permitirá visualizar la filogenia en otros programas como FigTree.



Para la visualización de datos accesorios (metadata) repetir el procedimiento, utilizando un archivo delimitado por tabulaciones como entrada e indicar “ID” al nombre de la columna con los nombres de las muestras.

4.6. Evaluación de la relación genética mediante polimorfismos de nucleótido único (SNPs)

La herramienta Snippy v4.4.5 (https://github.com/tseemann/snippy) se utilizará para el mapeo de los fragmentos de lectura de los aislamientos al genoma de referencia, y la identificación de SNPs (*Page*, 2016). La cepa de *Listeria monocytogenes* EGD-e, depositada en el Bio-proyecto PRJNA413700 (*National Center for Biotechnology Information, NCBI*) se utilizará como genoma de referencia.

4.6a Mapeo de los genomas a la referencia y llamado de SNPs

for f in *_1.fastq.gz; do snippy --cpus 16 --outdir snippy/${f%_1.fastq.gz} --ref lm_EGD_e.fna --R1 $f --R2 ${f%_1.fastq.gz}_2.fastq.gz; done

cd snippy

snippy-core --ref lm_EGD_e.fna BE*

snippy-clean_full_aln core.full.aln > clean.full.aln

Las secuencias obtenidas a través de diferentes mecanismos de recombinación pueden influir en la similitud de las secuencias de los genomas en un grado mucho mayor que las mutaciones puntuales heredadas verticalmente, las cuales representan la señal compartida de un ancestro común y permiten un mejor trazado de la evolución. La herramienta Gubbins (*Croucher*, 2015) se utilizará para el reconocimiento y remoción de sitios de recombinación.

conda activate

run_gubbins.py --threads 16 -p gubbins clean.full.aln

conda deactivate

snp-sites -c gubbins.filtered_polymorphic_sites.fasta -o clean.core.aln

4.6b Construcción de la filogenia

El alineamiento final clean.core.aln se utilizará para la construcción de la filogenia molecular utilizando IQ-tree (http://www.iqtree.org/), con un modelo de evolución GTR+I+G y estableciendo la confiabilidad de la topología mediante bootstrap con 10000 réplicas.

iqtree -s clean.core.aln -m GTR+I+G+ASC -pre gtr_global -bb 10000 -nt 16

4.6c Obtención de la matriz de SNPs de los aislamientos

La herramienta snp-dists (https://github.com/tseemann/snp-dists) permite obtener la diferencia de SNPs entre los aislamientos.

conda activate

snp-dists clean.core.aln > snp_matrix.tsv

conda deactivate

4.6d Visualización de la filogenia

La filogenia puede ser visualizada mediante el programa FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Para ello utilizar el archivo .treefile.

4.7. Análisis de genoma central o *core*

La herramienta Roary (*Page*, 2015) permite la identificación del pangenoma y genoma central. Para la obtención del pangenoma este pipeline agrupa los genes codificantes encontrados en la anotación de los genomas, de los diferentes aislamientos, de acuerdo con un porcentaje de similitud, definiendo un grupo de genes con mayor frecuencia (≥ 99%) denominado *core* y otros que tienen una menor frecuencia denominados como genes accesorios (soft, Shell y cloud).

4.7a Anotación de los genomas

Para la anotación se utilizará la herramienta Prokka (*Seemann*, 2014), la cual utiliza los ensamblados de *novo* de los aislamientos como archivo de entrada.

cd assemblies

for f in *.fasta; do prokka --prefix ${f%.fasta} --outdir prokka/${f%.fasta}.annotation --addgenes –genus Listeria --usegenus --strain ${f%.fasta} --mincontiglen 300 --cpus 16 $f; done

Obtención de los archivos .gff

cd prokka

for f in *.annotation/*.gff; do mv $f ${f%.annotation/*.gff}.gff; done

4.7b Obtención del pangenoma y genoma central, y llamado de SNPs

roary -e --mafft -f roary -p 16 *.gff

Llamado de SNPs

snp-sites -c core_gene_alignment.aln -o core_gene_alignment_snps.aln

4.7c Construcción de la filogenia

El alineamiento final se utilizará para la construcción de la filogenia molecular utilizando IQ-tree, con un modelo de evolución GTR+I+G y estableciendo la confiabilidad de la topología mediante bootstrap con 10000 réplicas.

iqtree -s core_gene_alignment_snps.aln -m GTR+I+G+ASC -pre gtr_pgl -bb 10000 -nt 32

4.7d Obtención de la matriz de SNPs de los aislamientos

conda activate

snp-dists core_gene_alignment_snps.aln > snp_matrix.tsv

conda deactivate

4.7e Visualización de la filogenia

La filogenia puede ser visualizada mediante el programa FigTree. Para ello utilizar el archivo .treefile.

4.8. Identificación de genes de virulencia

La herramienta ABRicate (https://github.com/tseemann/abricate) permite la identificación de genes de virulencia y resistencia a los antimicrobianos. Esta puede ser utilizada con múltiples bases de datos como: NCBI, CARD, ARG-ANNOT, Resfinder, MEGARES, EcOH, PlasmidFinder, Ecoli_VF and VFDB. Para la identificación de genes de virulencia se utilizará la base VFDB *core*.

cd assemblies

abricate --list

abricate --threads 8 --db vfdb /path/to/assemblies/*.fast > results.tab --mincov 99 --minid 95

El archivo results.tab contiene información detallada de los genes detectados para cada uno de los aislamientos, como por ejemplo el porcentaje de cobertura e identidad y localización en el ensamblado.

4.8a Obtención del resúmen de resultados

abricate --summary results.tab > summary.tab –identity

El archivo summary.tab contiene una matriz de presencia/ausencia de los genes. Un gen ausente es representado por un punto "." y un gen presente es indicado con el porcentaje de cobetura.

5.- BIBLIOGRAFIA

1. Alexey G, Vladislav S, Nikolay V, *et al*. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29: 1072-1075.

2. Antipov D, Korobeynikov A, McLean JS, *et al*. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 2016; 32:1009–1015.

3. Croucher NJ, Page AJ, Connor TR, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015; 43:e15.

4. Inouye, M., Dashnow, H., Raven, LA., *et al*. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014; 6:90.

5. Page AJ, Taylor B, Delaney AJ, *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016; 2:e000056.

6. Page AJ, Cummins CA, Hunt M, *et al*. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015; 31:3691-6393.

7. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30:2068-9.

8. Silva M, Machado MP, Silva DN, *et al*. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 2018; 4:000166.

9. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; 15:R46.

10. Zhou Z, Alikhan NF, Sergeant MJ, *et al*. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018; 28:1395-1404.